INTRODUCTION TO VLSI SYSTEMS

CARVER MEAD
Professor of Computer Science, Electrical Engineering, and Applied Physics, California Institute of Technology

LYNN CONWAY
Research Fellow, and
Manager, VLSI System Design Area
Palo Alto Research Center, Xerox Corporation

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California
London • Amsterdam • Don Mills, Ontario • Sydney

This book is in the
Addison-Wesley Series in Computer Science

Consulting Editor
Michael A. Harrison

Library of Congress Cataloging in Publication Data

Mead, Carver A
Introduction to VLSI systems.

1. Integrated circuits—Large scale integration.
2. Microcomputers. 3. Digital electronics.
4. Computer architecture. I. Conway, Lynn A., joint author. II. Title.
TK7874.M37 621.3819'535 78-74688
ISBN 0-201-04358-0

Second printing, October 1980

ISBN 0-201-04358-0
HIJKLMNOP-HA-8987654

1 MOS DEVICES AND CIRCUITS

We begin with a discussion of the basic properties of the *n*-channel, metal-oxide-semiconductor (MOS), field-effect transistor (FET). We then describe and analyze a number of circuits composed of interconnected MOS field-effect transistors. The circuits described are typical of those we will commonly use in the design of integrated systems. The analysis, though highly condensed, is conceptually correct and provides a basis for the solution of most system problems typically encountered.

Integrated systems in MOS technology contain three levels of conducting material separated by intervening layers of insulating material. Proceeding from top to bottom, the levels are termed *metal*, *polysilicon*, and *diffusion*, respectively. Patterns for paths on the three levels, and the locations of contact cuts through the insulating material to connect certain points between levels, are transferred into the levels during the fabrication process from *masks* similar to photographic negatives. (Details of the fabrication process will be discussed in Chapter 2.)

In the absence of contact cuts through the insulating material, paths on the metal level may cross over paths on either the polysilicon level or the diffusion level with no significant functional effect. However, wherever a path on the polysilicon level crosses a path on the diffusion level, a transistor is created. Such a transistor has the characteristics of a simple switch, with a voltage on the polysilicon-level path controlling the flow of current in the diffusion-level path. Circuits composed of such transistors, interconnected by patterned paths on the three levels, form our basic building blocks. With these basic circuits, we will design integrated systems, to be fabricated on the surface of monolithic crystalline chips of silicon.

1.1 THE MOS TRANSISTOR

An MOS transistor will be produced on the integrated system chip wherever a polysilicon path crosses a diffusion path, as shown in Fig. 1.1. The electrical sym-

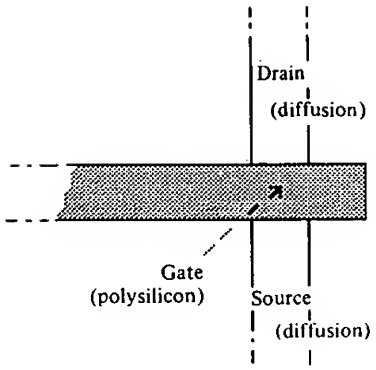

Fig. 1.1 MOS transistor, top view.

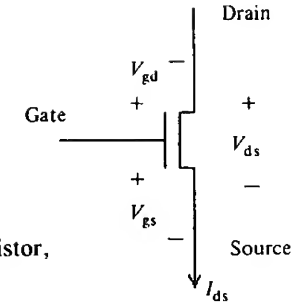

Fig. 1.2 MOS transistor symbol, subscripts in + to − direction sequence.

bol used to represent the MOS transistor in our circuit diagrams is shown in Fig. 1.2, along with symbols and polarities of certain voltages of interest. Note that the source and drain terminals of the device are physically symmetrical. For the *n*-channel MOSFET's the terminal labels are assigned such that drain-to-source voltage V_{ds} is normally positive. A more detailed view of the rectangular region called the gate, where the polysilicon (poly) crosses the diffusion, is given in Fig. 1.3. During fabrication the diffusion paths are formed after the poly paths are formed (as explained more fully in Chapter 2). The poly gate, and the thin layer of oxide beneath it, mask the region under the gate during diffusion. Therefore, no diffusion path forms under the gate, and there is no direct connection on the diffusion level between the source and drain terminals of the transistor. Metal, poly, and diffusion paths all conduct electricity well enough to be considered "wires" until further notice.

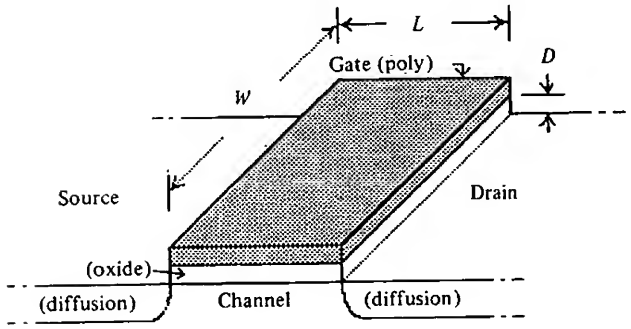

Fig. 1.3 MOSFET gate dimensions.

In the absence of any charge on the gate, the drain-to-source path through the transistor is like an open switch. The gate, separated from the substrate by the layer of thin oxide, forms a capacitor. If sufficient positive charge is placed on the gate so that gate-to-source voltage V_{gs} exceeds a *threshold voltage* V_{th}, electrons

will be attracted to the region under the gate to form a conducting path between drain and source. Most of the transistors we will use in our systems have threshold voltages greater than zero. These are called *enhancement mode* MOSFET's and their threshold voltage typically is ≈ 0.2 VDD, where VDD is the positive supply voltage for the particular technology.

The basic operation performed by the MOS transistor is to use charge on its gate to control the movement of negative charge between source and drain through the channel under the gate. The current from source to drain equals the charge induced in the channel divided by the *transit time* or average time required for an electron to move from source to drain. The transit time itself is the distance the electron has to move divided by its average velocity. In semiconductors under normal conditions, the velocity is proportional to the electric field driving the electrons. The relationship between drain-to-source current I_{ds}, drain-to-source voltage V_{ds}, and gate-to-source voltage V_{gs} is sketched in Fig. 1.4. For small V_{ds}, the transit time τ is given by Eq. 1–1:

$$\tau = \frac{L}{\text{velocity}} = \frac{L}{\mu E} = \frac{L^2}{\mu V_{ds}} . \tag{1–1}$$

The proportionality constant μ is called the *mobility* of the charge carriers, in this case electrons, under the influence of an electric field in the conducting material of the channel region. It is a velocity per unit electric field (cm²/volt-sec.). We shall see that the transit time is the fundamental time unit of the entire integrated system.

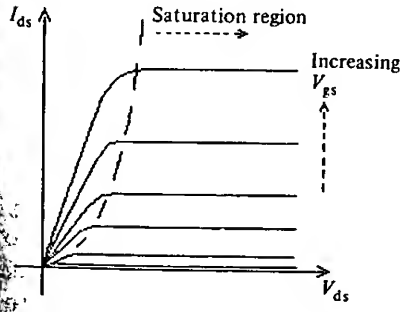

Fig. 1.4 Current versus voltage.

The amount of negative charge in transit Q is just the gate capacitance times the voltage on the gate in excess of the threshold voltage. The capacitance of two parallel conductors of area A, separated by insulating material of thickness D, equals $\epsilon A/D$. The proportionality constant ϵ is called the permittivity of the insulating material and has a simple interpretation. It is the capacitance of parallel conductors of area $A = 1$ cm², separated by a thickness $D = 1$ cm of the insulator material, and is in the units farad/cm. Therefore, the gate capacitance equals

$\epsilon WL/D$. Thus the charge in transit is given by Eq. (1–2), and the current is given by Eq. (1–3).

$$Q = -C_g(V_{gs} - V_{th}) = -\frac{\epsilon WL}{D}(V_{gs} - V_{th}) \tag{1–2}$$

$$I_{ds} = -I_{sd} = -\frac{\text{charge in transit}}{\text{transit time}} = \frac{\mu\epsilon W}{LD}(V_{gs} - V_{th})(V_{ds}) \tag{1–3}$$

Note that for small V_{ds}, the drain current is proportional to the source-drain voltage and also to the gate voltage above threshold. Any device with a current through it proportional to the voltage across it may be viewed as a resistor, and in the case of an MOS device with *low* drain-to-source voltage, the resistance is controlled by the gate voltage as given in Eq. (1–4).

$$\frac{V_{ds}}{I_{ds}} = R = \frac{L^2}{\mu C_g(V_{gs} - V_{th})} \tag{1–4}$$

In Eqs. (1–2) and (1–4), C_g is the gate-to-channel capacitance of the turned-on transistor. In the simple case where this transistor is driving the gate of an identical transistor, the time response of the system will be an exponential with a time constant RC_g, given in Eq. (1–5). This time constant is similar in form to the expression for transit time τ given in Eq. (1–1).

$$RC_g = \frac{L^2}{\mu(V_{gs} - V_{th})} \tag{1–5}$$

Although the above equations are greatly simplified, they provide sufficient information to make many design decisions that we will face, and they also give us insight to the scaling of devices to smaller sizes. In particular, the transit time τ can be viewed as the basic time unit of any system we shall build in the integrated technology. In almost all situations, the fastest operation that we can perform is to transfer a signal from the gate of one MOS transistor onto the gate of another. The transit time is the minimum time in which a charge placed on the gate of one transistor results in the transfer of a similar charge through that transistor's channel onto the gate of a subsequent transistor. For example, a transfer of charge from one transistor onto two identical transistors requires a minimum of two transit times. Thus, the transit time of the basic transistor in an integrated system can be viewed as the unit of time in which all other times in the system are scaled. Although it is a somewhat optimistic approximation, we will use τ as the primary time metric in calculating the delay through elementary inverting-logic stages. More accurate predictions of circuit behavior can be produced using any one of a number of available circuit simulation programs.[1,2]

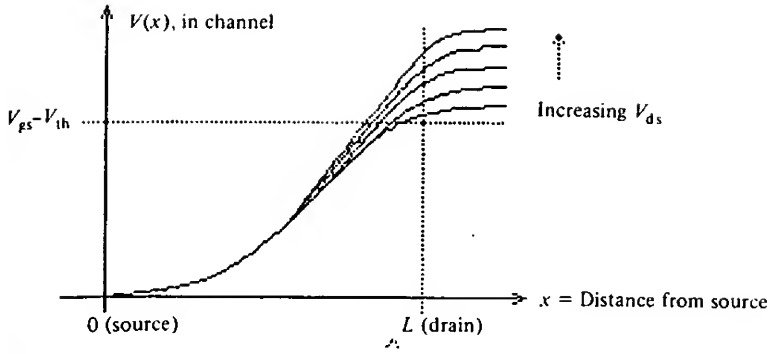

Fig. 1.5 Voltage profile across channel.

As V_{ds} is increased, not all the drain-to-source voltage is available for reducing the transit time. Drain voltage in excess of one threshold below the gate voltage creates a short region of high electric field, adjacent to the drain, that the carriers cross very quickly. The electric field in the major portion of the channel from the source up to this region is proportional to $V_{gs} - V_{th}$, as shown in Fig. 1.5. For $V_{ds} > (V_{gs} - V_{th})$, the drain current becomes independent of V_{ds}. Further increases in V_{ds} neither increase I_{ds} nor decrease the transit time. This range of V_{ds} values is known as *saturation*. In saturation,

$$I_{ds} = \frac{\mu \epsilon W}{2LD} (V_{gs} - V_{th})^2. \tag{1-6}$$

With the exception of the factor of 2 in the denominator, Eq. (1–6) is similar to Eq. (1–3), with the V_{ds} factor in (1–3) replaced by its maximum effective value, $V_{gs} - V_{th}$. The factor of 2 in Eq. (1–6) arises from the nonuniformity of the electric field in the channel region when in saturation.[3] (Richman, 1973)

1.2 THE BASIC INVERTER

The first logic circuit we will describe is the basic digital inverter. Analysis of this circuit is then extended to analysis of basic NAND and NOR logic gates. The inverter's logic function is to produce an output that is the complement of its input. When describing the logic function of circuits in integrated systems, we assign the value logic-1 to voltages equaling or exceeding some defined logic threshold voltage, and logic-0 to voltages less than this threshold voltage.

Were there an efficient way to implement resistors in the MOS technology, we could build a basic digital inverter circuit using the configuration of Fig. 1.6. Here, if the inverter input voltage V_{in} is less than the transistor threshold voltage V_{th}, then the transistor is switched off and V_{out} is "pulled up" to the positive supply voltage VDD. In this case the output is the complement of the input. If V_{in} is greater than V_{th}, the transistor is switched on and current flows from the VDD supply through the resistor R to GND. If R were sufficiently large, V_{out} could be

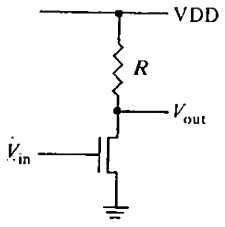

Fig. 1.6 An inverter.